# A Report on The Eleventh Language Testing Research Colloquium

**Achara Wongsothorn**

Chulalongkorn University Language Institute

This year (1989), San Antonio, Texas, the city of the Alamo and the famous American frontiers heroes—Jim Bowie, Davy Crockett, and the heroine, the Yellow Rose of Texas—witnessed two significant events which took place successively—the Eleventh Language Testing Research Colloquium (LTRC) and the Twenty-Third Annual Convention of the Association of the Teachers of English to Speakers of Other Languages (TESOL Festival' 89). The first event took place from March 4-6 at the legendary Emily Morgan Hotel while the second which ran from March 7-11 was held at the modern and spacious San Antonio Convention Center.

The Defense Language Institute (DLI) who hosted the 11th LTRC exhibited their interest in language aptitude testing by joining the panel discussion on this topic. The twenty papers presented can be classified into four themes as follows :

- Development and Standardization of Language Tests
- The Studies of the Components of Language Proficiency and Traits of Communicative Competence
- Uses of Language Tests for Language Teaching Issues and Problems in Language Aptitude Testing

**Theme I : Development and Standardization of Language Tests**

The three papers under this heading were :

1. Caroline Clapham. *ELTREV : Towards the Validation of an International EAP Test.*

2. James D. Brown. *Short-Cut Estimates of Criterion-Referenced Test Reliability.*

3. Carolyn Turner, *The Underlying Factor Structure of L2 Cloze Test Performance in Francophone, University-Level Students : Causal Modeling as an Approach to Construct Validation.*

From the titles we can see that the thrust of language test analyses was towards criterion-referenced test reliability and construct validity. Short-cut estimates of CRT reliability consisted of the threshold loss agreement approach, the squared-error loss agreement approach and the domain score dependability approach. Simplified calcula-

tions of data obtained from a single administration of the test were demonstrated as an alternative to the more elaborate and time-consuming pre-post test data method. The agreement and Kappa coefficients, the phi (lamda) dependability index and the phi coefficient could be obtained through simplified and short-cut statistics suggested by Brown.

The relevancy of the EAP demands of the subject areas to the test specifications is essential in maintaining test validity and avoiding the mismatch between the test and the actual text which will invalidate the whole testing process, voiced Clapham.

**Theme II : The Studies of the Components of Language Proficiency and the Traits of Communicative Competence**

There were eight papers presenting various aspects of the research which used language tests as tools for the analyses. They were :

1. Lyle Bachman, Fred Davidson, Brian Lynch and Katherine Ryan. *Content Analysis and Statistical Modeling of EFL Proficiency Test.*

2. Michael Milanovic. *The Construction and Validation of a Performance-Based Test Battery.*

3. Achara Wongsothorn. *Domain-Referenced Tests for Secondary and Tertiary Levels of Education.*

4. Gary Buck. *A Construct Validation Study of Listening and Reading Comprehension.*

5. Kyle Perkins and Charles Parrish. *The Determination of Hierachies among TOEFL Reading Comprehension Items.*

6. Dan Douglas. *Strategic Competence and the SPEAK Test : An Exploration of Construct Validity.*

7. Liz Hamp-Lyons and Sheila Prochow. *Person Dimensionality, Person Ability and Item Difficulty in Writing.*

8. Grant Henning. *Effects of Short Term Memory Load, Reading Response Length, and Processing Hierachy on TOEFL Listening Comprehension Item Performance.*

Bachman presented the results of the content analyses of the reading sections the Test of English as a Foreign Language (TOEFL) and the Certificate of Proficiency in English (CPE) or First Certificate in English (FCE), together with results of analyses of the testees' performance. Bachman used the communicative and test-method framework in which all factors influencing students' test scores were delineated to analyse the comparability of the content of the two tests. These factors are composed of language abilities and methods of testing. Discussion of the relevance of both the *a priori* analysis of testees' responses for test development and the use of language tests ensued.

In Milanovic's study, the validation of a model of communicative competence through task level factor analysis supported the notion that communicative competence consists of two separate traits : listening and reading-writing, as well as illustrating the presence of grammatical and socio-linguistic skills at all levels of ability. At a high level of language proficiency, the traits appear to be more flexible and not as stable as at low levels of proficiency. It appears that the factor structure is more clear-cut at a lower level of proficiency.

Along the same line of research, Buck found that listening and reading comprehension are two separate but closely correlated traits.

In a study with Thai secondary and tertiary students, Wongsothorn found that grammar, vocabulary and phonology shared common variances with sound and graphic modalities indicating that they could be subsumed under communicative modalities. The research also revealed that the number and traits of language learning factors differed among levels with lower secondary level having more sound modality factors while the tertiary level had more graphic modality factors. SSQ analysis illustrated that sound modality shared greater variances with domain-referenced proficiency at the ratio of 1.8 : 1, and that both modalities had common variance with domain-referenced proficiency by approximately 50%.

Kyle Perkins and Charles Parrish's research aimed to identify and describe the significant prerequisite relations which underlie item responses from one form of TOEFL vocabulary and reading comprehension test, They analysed items' frequency of occurrence and dispersion traits for vocabulary, and text structure, facet theory and comprehension category framework for reading comprehension. Factor analysis procedures indicated that only one content factor was present, thereby nullifying the notion of multiple prerequisite relationships.

Douglas analysed transcribed protocols of authentic spoken texts by non-native and native speakers of English on the spoken part of TOEFL. He studied the associations between planning and execution strategies of the testees and their scores through the examination of the grammatical, textual, illocutionary and sociolinguistic data. In analyzing the testees talk about the best way to ease a food shortage, the uncovered differences in planning and execution strategies between native and non-native speakers of English noticing that repetitions and false starts had the greatest degree of variance between groups.

Questioning the use of single holistic scoring in grading essays, Liz Hamp-Lyons studied relationships among the performance (improvement), an index of frequency of scores, holistic scores, language proficiency level (person ability) and item difficulty. Hamp-Lyons concluded that holistic scoring might not be appropriate because writing is a multifaceted rather than unidimensional process. Further research into psychometrical properties of writing ability is suggested by the researcher.

Henning's study on TOEFL listening comprehension though multitrait-multimethod validation pinpointed superiority of item format measured by effects of levels of repetition, passage length, reading task length, and hierachy of response processing on item difficulty and discriminability. The study suggested that listening test items which directly measure listening, not reading nor other types of deciphering (code-breaking) skills, would be more valid in assessing students' listening skills.

**Theme III : Uses of Language Tests for Language Testing**

Papers classified into this thematic heading consisted of :

1. Margaret Des Brisay. *The Problem of the Middle Ground. Where Do You Draw the Line ?*

2. Doreen Ready and Robert Courchêne. *The Use of Indirect Measures to Test L1/L2 Writing Ability. Some Issues.*

3. James D. Brown. *The Place of ESL Students in Writing Across the Curriculum Placement Test Battery.*

4. Kathy Baily, Peter Shaw and David Tsugawa. *Assessment Implications of a Content-Based Curriculum : The Role of Self- Assessment.*

5. Neil Anderson. *Taking Reading Comprehension Tests : What Are Second Language Readers Doing?*

6. Andrew Cohen. *The Taking and Rating of Summary Tasks.*

The problem of the validity of cut-off points in selecting students and eliminating them was regarded as unfair danger in language education. Des Brisay used the quotation from Robert Frost to illustrate her concern :

*Before I'd build a wall,*
*I'd ask to know.*
*Who I was walling in or walling out.*

The role of self-assessment in a content-based curriculum discussed in the light of its criterion-related validity, its practicality and its benefits to the language learners led to the following suggestions by Baily, Shaw and Tugawa :

(1) self-assessmênt could provide one corpus of information for quantitative-qualitative schemes of student evaluation,

(2) self-assessment instruments for rating students' own knowledge of their subject field could be useful for the comparison of students' mastery of subject matter in content-based courses with that in their regular courses taught in English,

(3) teachers could obtain a good estimate of the students' communicative ability from results of students' self-assessed data rather than their linguistic accuracy because self-assessment instruments were based on communicative principles of language use,

(4) self-assessment instruments could provide information about students' communicative proficiency and also an indication of their confidence in rating their own level of language ability,

(5) with the use of students' ethnographic data, self-assessment instruments could provide diagnostic information about students' language problems as well as predictive information about their success in the subject fields.

In Carolyn Turner's research, a causal modeling approach to construct validation was used to investigate the underlying factor structure of students' performance on an L2 cloze test. The research uncovered nonlinguistic factors which contributed to students' achievement, besides their L1 and L2 proficiency.

In placement testing of students' writing, Brown studied groups of native speakers of English and ESL students the University of Hawaii at Manoa. The concern for student general ESL needs should be coupled with their EAP needs.

Courchêne and Ready investigated the effectiveness of indirect measures to test L1/L2 writing ability. The study indicated that an indirect measure test is not adequate in placing students nor in providing the diagnostic information that direct measures can.

Cohen's study on summary tasks focused on the summarizer and the rater. The text

and the context in which the test is conducted were also regarded as essential. It is essential that the summarizers know the following :

(1) number of words/sentences expected

(2) allocation and reduction of points

(3) whether a first draft or a finished product is expected

(4) whether the summary is expected to be a listing of main points or a logical integration of main points?

For the raters, information about the following points should be provided :

(1) How to score summary papers—whether the raters are to read all respondents summaries of one text and then move on to the summaries of the next text, or rate each paper one after another.

(2) whether a rigorous rating scale is provided together with frequent checks to ensure inter-rater reliability.

(3) whether provision is made to ensure that raters who are native speakers of the original foreign-language source text participate in the rating together with native speakers of the summarized texts to ensure the accuracy of interpretation.

Instructions for summarizers and the techniques of summarizing—parallel (same or similar elements), sparallel-selective (one or several elements inserted), selective (selected from scattered sequences), selective-generalizing (generalization or synthesis of elements), integrating (generalizing using specific items) and commentary (expressing something not expressed in the source text)—should also be analyzed as well as text characteristics.

In studying the strategies that L2 students with different levels of language proficiency (beginning, intermediate and advanced) employed while taking a standardized reading comprehension test, Neil Anderson administered the Textbook Reading Profile (TRP) —a self-assessment checklist—to all groups. Think-aloud protocols made in Spanish and English by students describing their reading and testing strategies were recorded. Anderson discovered that advanced students could ask themselves questions about what they were reading and finding in the text. They could also perform better than lower proficiency students in making judgement about their ability towards the text, in responding effectively to the text, in speculating beyond information in the text, and in making guesses, etc. Lower level proficiency students, on the other hand, were more concerned with time limit for test taking, and did not read the test questions before beginning the reading as often, compared to advanced students. From the 47 processing strategies studied, seven are given here : (1) developing awareness, (2) accepting ambiguity, (3) establishing lexical ties, (4) establishing intrasentential ties, (5) establishing intersential ties, (6) using background knowledge and (7) using test taking skills. The research tended to provide information on how students actually take reading comprehension tests as opposed to what they think they were expected to be doing.

**Theme IV : Development of Language Aptitude Tests**

The three papers in this theme were :

(1) Charles Stanfield. *A Rationale for a Reexamination of Language Aptitude Testing.*

(2) Thomas Parry and James Child. *Preliminary Investigation of the Relationship between VORD, MLAT and Language Proficiency.*

(3) John Lett. *Predictors of Success in an Intensive Language Learning Context: An Explanatory Model of Classroom Learning.*

Language aptitude tests such as Pimsler PLAB, Carroll and Sapon MLAT, and the DLI ALAT have been used extensively for administrative and academic purposes in language program planning and language instruction. Stanfield emphasized the necessity of reexamining the validity of the existing measures as well as the newly developed ones.

The study of psychometric properties of the factor called language aptitude using VORD by the DLI and the CIA proved that language aptitude is more than a unidimensional construct and that MLAT is more effective than VORD in predicting language learning achievement. However, research findings indicated that VORD might be a better predictor of achievement in discrete linguistic knowledge rather than language communication.

Data from 1903 to 1987 were analyzed for the joint project of the DLI and the US Army Research Institute (ARI) using selected army enlisted personnel learning Korean, Russian, German, and Spanish as the subjects of study. Lett, Jr. reported that speaking skills are less predictable than listening and reading and that for difficult languages like Korean and Russian, aptitude has influential effect. General ability appeared to predict achievement of less difficult languages like Spanish and German. Both cognitive and non-cognitive variables are important and may not operate in the same way for all languages, Lett stated.

The general atmosphere of the 11 th LTRC was densely academic but interspersed with light friendly events—the breaks, morning coffee and the semi-formal buffet party which was meant to be a welcoming reception—or possibly also a get-together of language testing researchers from both Europe, Asia and America. Most if not all of the participants continued their stay to include the 23 rd TESOL Convention and attain the joy of strolling along the River Walk or even of cruising along the winding green Santonio River.